COMP 4980 “Introduction to NLP”

Text summarization

Stan Szpakowicz

Emeritus Professor

School of Electrical Engineering & Computer Science, University of Ottawa

TRU, Fall 2017

Dan Jurafsky’s slides, reshaped by Stan Szpakowicz

The plan

Summarization terminology

The goal is rather obvious

A summary is an abridged version of a text which contains information important or relevant to a user. Note how conveniently relative it is! Different users have different needs.

So, a summary can be:

- an outline or an abstract of a document such as an article, a manual, an opinion piece, *etc.*;
- a recap of an email thread;
- a list of action items from a meeting;
- a text simplified by compressing sentences;
- and so on, and so on.

Single and multiple documents

Single-document summarization, as the name suggests, condenses a document.

Multiple-document summarization can work on:

- a series of news stories on the same event — maybe an evolving story;
- a set of Web pages, the result of a search on some topic or question;
- and so on.

Rather predictably, multiple-document summarization is a good deal more difficult, though even a non-trivial single document can be a big challenge.*

*To summarize a long novel *well*, for example, is a dream.

Another dimension: focus

Generic summarization produces the abridged content to serve perhaps many purposes.

One very simple idea: find keywords in a text,* and gather some of the sentences in which those keywords occur.

Query-focused summarization is guided by an information need expressed in a user query.

> It is actually a form of complex question-answering:
>
> answer a question by summarizing a document which contains the information we require.

The focus need not be a query. One can, for example, try to summarize a story by concentrating only on person, location and time mentions in the text.

---

* *Keyword extraction* is a separate NLP task, less complicated than summarization, though not easy either.

Google summarizes, sort of

Dan Jurafsky's choice for this little demo is Die Brücke.

Every hit has a title, a link and a *snippet*. It is all you get to help you decide which hit, if any, is worth a click.

Clearly, it would be a lot more useful if Google created a coherent answer which combines information from each document — without opening any of them. Now that would be multiple-document summarization worth having!

Snippets seem to be too trivially acquired to be much good. They come from the opening bit of a page.*

To be fair, Google snippets sure are query-focused.

*This, as it happens, works quite well for news. Good journalistic style dictates that a news item begins with — yup! — a summary of what follows.

One more dimension

Extractive summarization:

> create the summary from phrases or sentences in the source document(s).

Abstractive summarization:

> express the ideas in the source documents using (at least in part) different words.

Abstractive summarization would be much more useful, were it not much too difficult to pull off.

We have to live with the drawbacks of extractive summarization, such as lack of continuity, lack of cohesion, and dangling references.†

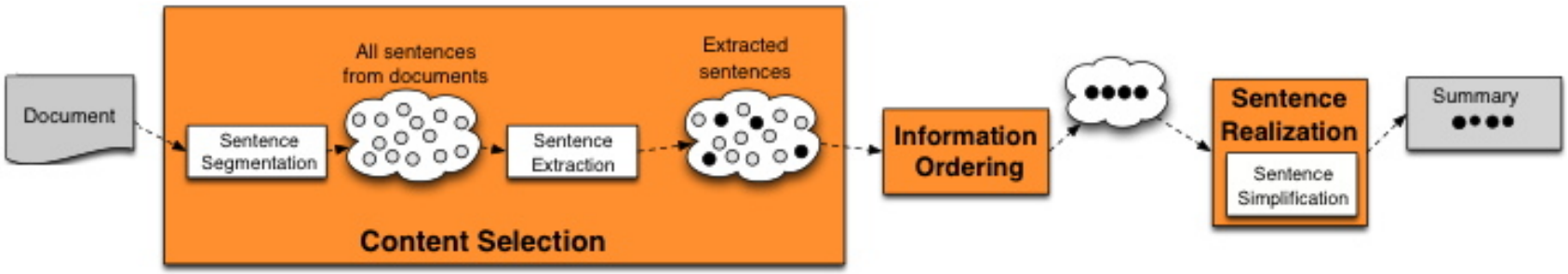
† That would be, *e.g.,* a pronoun whose *antecedent* did not make it into the summary.

The plan

. . . a quick overview

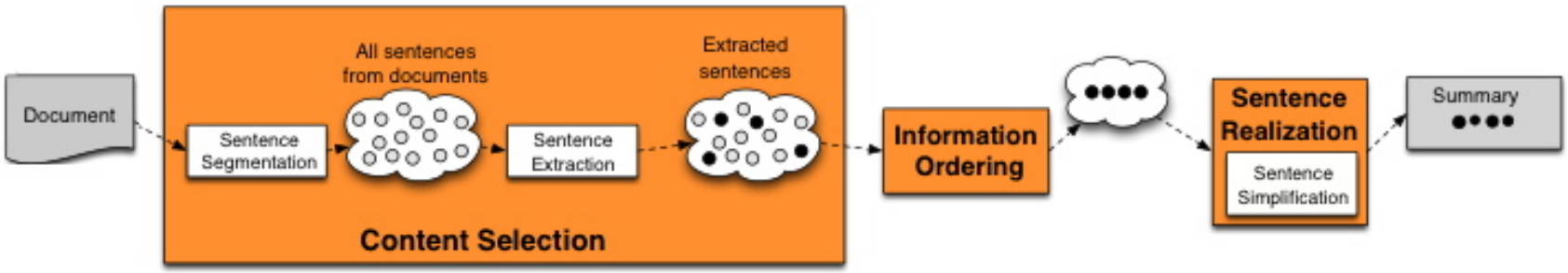
Three stages

1. Content selection:
   choose sentences to extract from the document.
2. Information ordering:
   choose an order in which to place them in the summary.
3. Sentence realization:
   clean up the sentences (if necessary).



Three stages — the basic version

1. Content selection:
   choose sentences to extract from the document.
2. Information ordering:
   use document order.
3. Sentence realization:
   keep the original sentences.



Unsupervised content selection

The intuition for content selection in extractive summarization goes back to all the way Luhn 1958:

> sentences with salient words are best for a summary.

1. We score all *content* words.
2. A score above a predetermined threshold puts a word in a topic signature.
3. We select sentences rich in topic-signature words.

Scoring is key.

In constructing topic signatures, we usually rely on tf-idf or on the log-likelihood ratio.

tf-idf in summarization

We might simply count content words — filtering out stop words, the more the better.*

That would not do if we wanted a decent summary. We would assume — unrealistically — that words are all independent. That, as we well know, is not true.

Generally speaking, tf-idf finds salient words: those with many occurrences in few documents. Such words characterize those documents. *Words present in all or in most documents are not good discriminants, never mind whether they are very frequent closed-class words or not so frequent open-class words.*

In one document, sentences with many salient words are also likely to be somehow characteristic of that document's topic.

*For example, *light* verbs such as get, set, put, let, *etc.* are not informative.

The log-likelihood ratio

LLR a rather complicated statistic, which people tend to use without knowing quite what it does.* The concept is well-known in medicine:

> The Likelihood Ratio (LR) is the likelihood that a given test result would be expected in a patient with the target disorder compared to the likelihood that that same result would be expected in a patient without the target disorder.†

Here, we consider salient words in a document and the same words in general language. Take "method", ranked 1001 on a free list of top 5000 words. In a 15000-word document, we expect ∼15 occurrences, but suppose we get 40. The surplus of 25 is what makes "method" salient in this document.

---

*Look at Dunning's paper is you feel up to it.

†I borrowed this definition from the CEBM site.

The log-likelihood ratio (2)

The actual formula people use in summarization looks a little like magic. (The likelihood ratio is typically denoted as λ.)

Someone clever determined that $\boxed{-2\,log\,\lambda(w) > 10.83}$ is all it takes for word *w* to be statistically *highly significantly* more frequent in a document than in general English.*

People are lazy, so they cut the fraction ☺:

$$weight(w) = \begin{cases} 1 & \text{if -2 } log\,\lambda(w) > 10 \\ 0 & \text{otherwise} \end{cases}$$

Put differently, salient words contribute, other words do not.

*That is the significance level α = 0.001, or 1‰ chance of being wrong.

LLR and query focus

A variation on the pure LLR-based construction of topic signatures is the addition of words which are informative because they appear in a query.

$$weight(w) = \begin{cases} 1 & \text{if } -2\, log\, \lambda(w) > 10 \\ 1 & \text{if } w \in query \\ 0 & \text{otherwise} \end{cases}$$

Finally, here is how we calculate sentence weight:

$$weight(S) = \frac{1}{|S|} \sum_{w \in S} weight(w)$$

Supervised content selection

We can do that if we happen to have a labelled training set of good summaries for each document.

1. Align document sentences with sentences in the summary.
2. Extract features:
   - sentence position in the document, *e.g.,* first;
   - sentence length;
   - word informativeness. And so on.
3. Train a binary classifier (summary-worthy sentence: yes/no).

But:

- it is hard to get labelled training data,
- alignment is difficult,
- performance is on par with unsupervised algorithms.

In practice, unsupervised content selection is more common.

The plan

. . . rounding it off

Easy summaries are not good summaries

It really is trivial to produce a summary. To construct an automatic summary of sufficient quality is another matter.

It is also not clear what makes a summary good..

People have devoted a good deal of attention to the evaluation of summaries. A fine place to start is the DUC site.

On small NLP tasks,* human evaluation is almost always more useful than automated evaluation. It is quite costly, however, and annoyingly uneven.

Naturally, the research community seeks automation even if the quality of evaluation — the feedback for summarizing system designers — is often low.

*Summarization on a scale practiced thus far was a small task.

ROUGE

Recall-**O**riented **U**nderstudy for **G**isting **E**valuation

Intrinsic metric for evaluating summaries quickly and cheaply:

- based on BLEU (a metric used in machine translation);*
- not nearly as good as human evaluation ("Did the summary answer the user's question?", "Is it relevant?", *etc.*);
- but much more convenient.

Given a document:

1. have a number of people (more is better) write the document's reference summaries $\mathcal{R}$;
2. have the automatic summarizer make a summary A;
3. get the % of the bigrams in $\mathcal{R}$ which also appear in A.

*Ask Wikipedia about **Bi**Lingual **E**valuation **U**nderstudy. And guess why the acronym **ROUGE**...

ROUGE (2)

Here is the formula for one of several metrics which ROUGE offers. It is named *ROUGE-2*. $\mathcal{R}$ is the set of reference summaries, A is the automatic summary, b is a bigram in a summary.

$$\frac{\sum_{s \in \mathcal{R}} \sum_{b \in s} min(count(b, A), count(b, s))}{\sum_{s \in \mathcal{R}} \sum_{b \in s} count(b, s)}$$

There also are *ROUGE-1*, *ROUGE-3* and possibly more.

ROUGE-L matches sequences of words using LCS.

ROUGE-S matches skip-grams: word pairs that can have a maximum of N gaps between words. We have, *e.g.,* skip-bigrams.

For example, in cat in the hat the skip-bigrams would be cat in, cat the, cat hat, in the, in hat, the hat.

ROUGE, a toy example

Reference 1

Water spinach is a green leafy vegetable grown in the tropics.

Reference 2

Water spinach is a semi-aquatic tropical plant grown as a vegetable.

Reference 3

Water spinach is a commonly eaten leaf vegetable of Asia.

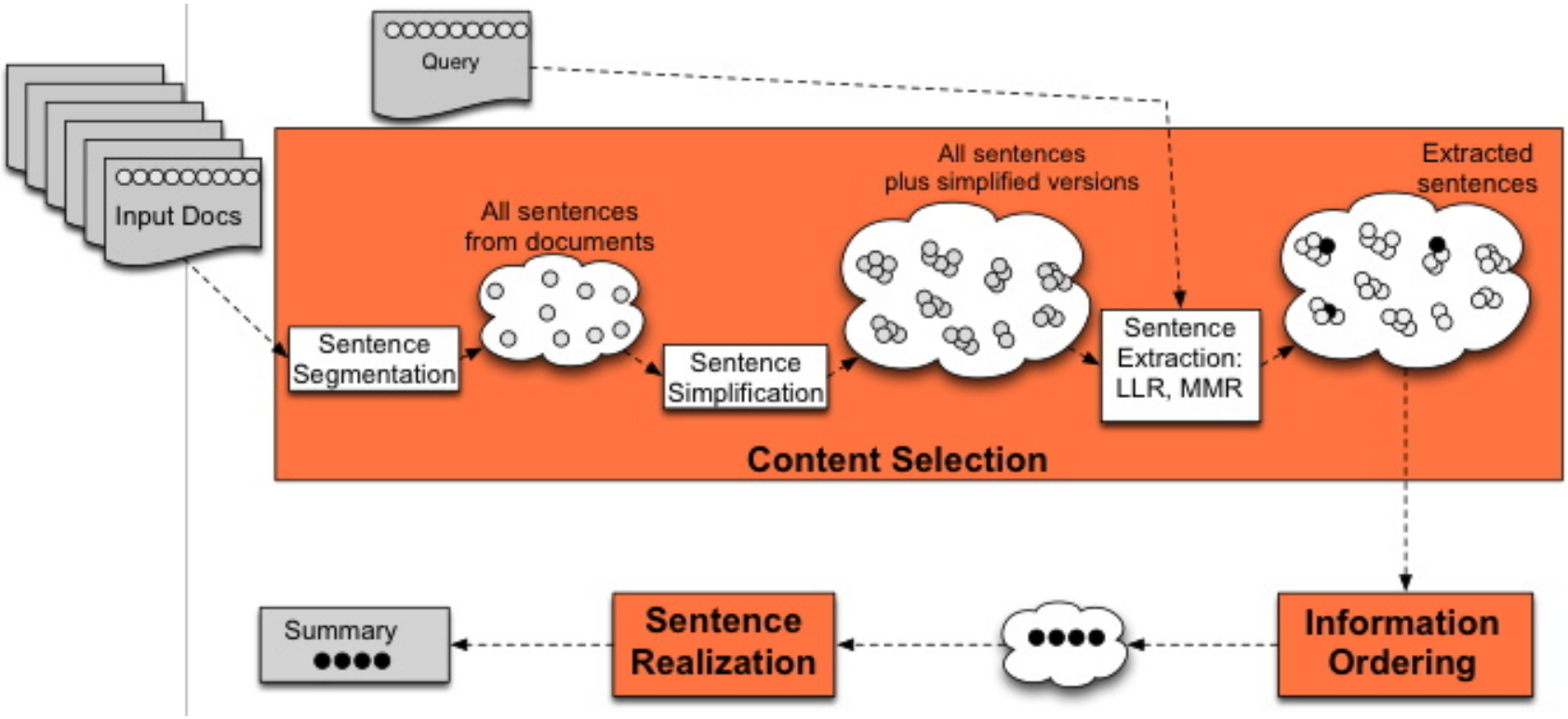
Automatic summary

Water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.

$$\textit{ROUGE-2}(\textit{these sentences}) = \frac{3+3+6}{10+10+9} \approx 0.4138$$

This number is hugely inflated. In DUC evaluations, most ROUGE values were something like 0.01 to 0.04 ☹.

Query-focused multi-document summarization



Sentence simplification — remove less important pieces

Parse sentences; use rules to decide which modifiers to prune.

appositives	Rajam ~~, 28, an artist who was living at the time in Philadelphia,~~ found the inspiration in the back of city magazines.
attribution clauses	Rebels agreed to talks with government officials~~, international observers said Tuesday~~.
PPs without named entities	The commercial fishing restrictions in Washington will not be lifted unless the salmon population increases ~~to a sustainable number~~.
initial adverbials	~~For example,~~ ~~On the other hand,~~ ~~As a matter of fact,~~ ~~At this point,~~

and so on.

Maximal marginal relevance

Iterative content selection from multiple documents (1998)

- Greedily choose the best sentence to insert in the growing summary.

  The next sentence will be relevant and novel:

  - maximally relevant to the user's query — high similarity to the query (*e.g.*, cosine);
  - minimally redundant with the summary so far — low similarity to the summary (*e.g.*, cosine).

  $$\mathit{argmax}_{c_i \in R \setminus S} (\lambda \mathit{sim}_1(c_i, Q) - (1-\lambda) \max_{c_j \in S} \mathit{sim}_2(c_i, c_j))$$

- Stop when the desired length has been reached.

R: relevant documents; S: summary; Q: query; λ: a parameter; sim_k: similarity measure; c_k: candidate sentence

LLR and MMR

How to choose informative yet non-redundant sentences?

The intuitions of LLR and MMR can be combined in many useful ways.

Here is one possibility.

- Score each sentence using LLR. Include query words in the calculations.
- Put the sentence with the highest score in the summary.
- Iteratively add high-scoring sentences which are not redundant with the summary so far.

Information ordering

Chronological ordering:

order sentences by the date of the document (for news summarization).

Ordering by coherence:

choose orderings

which make neighbouring sentences similar;

in which neighbouring sentences discuss the same entity.

Topical ordering:

learn the ordering of topics in the source documents.

Back to earth

All that was quite advanced (though I practiced evasion on the last two slides ☺).

For the demo, we have two *very* simple programs. Neither of them applies any of the more elaborate scoring / measuring techniques.

Go visit the course Web site, the page with the class notes.

The plan

That's it

Well, nearly it. . .

. . . expect one more set of class notes. ☺